

# Energy Management in Heterogeneous Networks with Cell Activation, User Association and Interference Coordination

Quan Kuang, and Wolfgang Utschick, *Senior Member, IEEE*

**Abstract**—The densification and expansion of wireless network pose new challenges on interference management and reducing energy consumption. This paper studies energy-efficient resource management in heterogeneous networks by jointly optimizing cell activation, user association and multicell multiuser channel assignment, according to the long-term average traffic and channel conditions. The proposed framework is built on characterizing the interference coupling by pre-defined interference patterns, and performing resource allocation among these patterns. In this way, the interference fluctuation caused by (de)activating cells is explicitly taken into account when calculating the user achievable rates. A tailored algorithm is developed to solve the formulated problem in the dual domain by exploiting the problem structure, which gives a significant complexity saving. Numerical results show a huge improvement in energy saving achieved by the proposed scheme. The user association derived from the proposed joint resource optimization is mapped to standard-compliant cell selection biasing. This mapping reveals that the cell-specific biasing for energy saving is quite different from that for load balancing investigated in the literature.

**Index Terms**—cell activation, power minimization, resource management, user association, interference coupling, channel allocation, range expansion, cell selection biasing, interference coordination, cutting plane methods.

## I. INTRODUCTION

There exists an emerging paradigm shift in wireless infrastructure systems, where densely deployed small and low-cost base stations (BSs) are embedded into the conventional cellular network topology to form a so-called heterogeneous network (HetNet) [1]. In a dense HetNet, BSs are typically deployed to satisfy the peak traffic volume and they are expected to have low activity outside rush hours such as nighttime. Hence, there is a high potential for energy saving if BSs can be switched off according to the traffic load [2].

Obviously, cell activation is coupled with user association: the users in the muted cells must be re-associated with other BSs. In addition, cell muting and user re-association impose further challenges on interference management, since the user may not be connected to the BS with the strongest signal

strength. This interference issue can be resolved by resource coordination, i.e., properly sharing the channels among multiple cells and then distributing them to the associated users in each cell. Hence, to achieve energy efficiency, multicell multiuser channel assignment should be integrated into the optimization of the cell activation and user association.

However, the resource management that considers the above elements jointly is very challenging mathematically because the inter-cell interference coupling leads to the inherent non-convexity in the optimization problems. To make the problems tractable, the previous studies relied on worst-case interference assumption [3], [4], average interference assumption [2], [5], or neglecting inter-cell interference [6]. In these works, the interference was assumed *static* (or absent), i.e., independent of the resource allocation decisions in each cell when estimating the user achievable rate. Clearly, this is a suboptimal design because the BS deactivation will cause interference fluctuation in the network, hence affecting the user rate.

This paper develops a new framework for energy-efficient resource management to consider the coupling effect of the inter-cell interference caused by cell activation. The idea is to pre-calculate the user rate under each possible *interference pattern* (i.e. an interference scenario in the network, described as one combination of ON/OFF activities of the BSs), and then perform resource allocation among these patterns. This allocation yields the actual interference and the corresponding user achievable rates that well match the interference at the same time.

Other related works include [7], [8], [27] and references therein, where the sparse optimization techniques similar to one adopted in this paper have been used to optimize BS activation and/or coordinated multi-point processing (CoMP) according to the instantaneous channel state condition, with pre-determined channel allocation. By contrast, we consider a slow adaptive strategy over a period of many minutes to reduce the control overhead. Moreover, channel allocation is introduced as an effective way for interference mitigation.

The system model and problem formulation in this work stem from our prior studies in [9], [24], where multi-pattern resource allocation were exploited to maximize the network proportional fairness utility with a fixed number of active cells. The similar idea of allocating spectrum among multiple reuse patterns in a slow timescale has also been adopted in independent works [10], [26] for optimizing queueing delay and/or energy management. Nevertheless, in this paper a unique dual cutting plane approach is used to develop efficient

This is an extended version of a paper to appear in IEEE Transactions on Wireless Communications (TWC) with the same title, containing all detailed proofs. Manuscript submitted to IEEE TWC July 15, 2015; revised November 25, 2015; accepted January 29, 2016 Part of the work has been accepted to present at ICASSP 2016 [28].

Q. Kuang was with the Department of Electrical and Computer Engineering, Technische Universität München, Munich 80333, Germany. He is now with Panasonic R&D Center Germany (e-mail: quan.kuang@eu.panasonic.com). W. Utschick is with the Department of Electrical and Computer Engineering, Technische Universität München, Munich 80333, Germany.

solving algorithms (see a summary of contributions below).

The main contributions of this paper are:

- 1) An energy-efficient resource management problem is proposed to serve the user demand with minimum network power consumption. It jointly optimizes the cell activation, user association, and multicell multiuser channel assignment, according to the long-term average traffic and channel conditions. The interference coupling caused by cell activation is taken into account when estimating the user achievable rates, which is done by pre-defining interference patterns. Although a network of  $B$  cells has  $2^B$  possible interference patterns, we show that a small number of patterns are sufficient to obtain accurate estimates of the user rates.
- 2) A tailored algorithm is developed for solving the formulated resource management problem. There are two ingredients in this development: the reweighted  $\ell_1$  minimization [11] is used to tackle the  $\ell_0$  term in the objective, and a cutting plane method is used for solving the dual problem by exploiting the problem structure, resulting in a significant complexity saving compared to directly applying standard interior-point solvers. This complexity reduction makes it possible to include all  $2^B$  patterns in the optimization problem for reasonably-sized networks, providing an energy saving benchmark for comparison with other schemes where pattern selection is restricted.
- 3) Using the proposed framework, existing resource management proposals are evaluated and compared in a unified manner, where the interference coordination is either involved or not, and the user association is either optimized jointly with the resource allocation or performed by simple cell selection biasing. In this way, the impacts of interference coordination and user association on energy saving are individually characterized.
- 4) The user association decision obtained by the proposed joint resource optimization is mapped to standard-compliant cell selection biasing [12]. This mapping reveals that, in contrast to previous studies (e.g., [13]) where common per-tier biasing is sufficient in terms of load balancing and network rate utility maximization, the energy-efficient solution requires individual biasing for different cells in the same tier.

The remainder of this paper is organized as follows. Section II introduces the system model. Rate-constrained energy saving problem is formulated and studied in Section III. Section IV develops a unified view on a wide range of existing resource management schemes, and the performance comparison is provided in Section V followed by conclusion in Section VI.

## II. SYSTEM MODEL

We consider a downlink HetNet, where a number of small cells<sup>1</sup> are embedded in the conventional macro cellular network. The set of all (macro and small) cells is denoted by  $\mathcal{B} = \{1, 2, \dots, B\}$ . The cells can be switched on or off

every time period  $T$  (say, many minutes) according to the fluctuations in the traffic load. We select  $K$  test points in the network as the representation of typical user locations [4], denoted by  $\mathcal{K} = \{1, 2, \dots, K\}$ . The user demand of each test point  $k \in \mathcal{K}$  is represented by a minimum required average rate  $d_k$ , which is assumed known from traffic estimation and user QoS requirements (see Section V-E for more discussion on test point selection and demand modeling). We are interested in developing adaptive strategies for every period of  $T$  to accommodate the user demand with minimum network energy consumption, taking into account the inter-cell interference coupling.

The enabling mechanism is to characterize the interference by specifying the interference patterns, each of which defines a particular ON/OFF combination of BSs. We use the pattern activity vector  $\mathbf{a}_i = (a_{i1}, a_{i2}, \dots, a_{iB})^T$  to indicate the ON/OFF activity of the BSs under pattern  $i$ , where

$$a_{ib} = \begin{cases} 1 & \text{if BS } b \text{ is ON under pattern } i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

We denote the set of pre-defined patterns by  $\mathcal{I} = \{1, 2, \dots, I\}$  and further define the matrix  $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_I)$  to combine the activity vectors for all candidate patterns. In order to fully characterize the interference scenarios in a network of  $B$  cells, generally speaking,  $2^B$  patterns are needed. However, since BSs located far away have weak mutual interference, omitting some patterns will not affect the accurate estimation of user achievable rates. We will discuss more on this next (see Proposition 1 and Section IV-C).

Fig. 1 illustrates the idea of multi-pattern formulation. Firstly, the multi-cell channel allocation is translated into partitioning the spectrum across all patterns. In a slow timescale considered in this paper, all frequency resources can be assumed to have equal channel conditions. Denote the spectrum allocation profile by  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_i, \dots, \pi_I)^T \in \Pi$ , where  $\pi_i$  represents the fraction of the total bandwidth allocated to pattern  $i$  and  $\Pi = \{\boldsymbol{\pi} : \sum_{i \in \mathcal{I}} \pi_i = 1, \pi_i \geq 0, \forall i\}$ . Then the total bandwidth fraction allocated to BS  $b$  is  $\mathbf{A}_{b\bullet} \times \boldsymbol{\pi}$ , where  $\mathbf{A}_{b\bullet}$  is the  $b$ -th row of the matrix  $\mathbf{A}$ .

Secondly, test point association and multiuser channel allocation can also be easily done thanks to the multi-pattern formulation. In more detail, denote by  $\alpha_{kbi} \geq 0$  the fraction of resources that BS  $b$  allocates to test point  $k$  under pattern  $i$ . Naturally, each BS is allowed to use up to  $\pi_i$  resources under pattern  $i$  for its associated test points, expressed as  $\sum_{k \in \mathcal{K}} \alpha_{kbi} \leq \pi_i, \forall b, \forall i$ . Note that the association is implicitly indicated by  $\alpha_{kbi}$ , i.e.,  $\alpha_{kbi} > 0$  means test point  $k$  is associated with BS  $b$  under pattern  $i$ , while zero value of  $\alpha_{kbi}$  means that they are not connected.

Finally, we define the usage of BS  $b$  as  $\rho_b = \sum_k \sum_i \alpha_{kbi}$ , with  $0 \leq \rho_b \leq 1, \forall b \in \mathcal{B}$ . Then cell  $b$  is active if  $\rho_b$  is nonzero.

### A. Rate model

Assuming flat power spectral density (PSD) of BS transmit power and the noise, the received SINR of the link connecting BS  $b$  to test point  $k$  under pattern  $i$  is

$$\text{SINR}_{kbi} = \frac{a_{ib} P_b G_{bk} \|h_{bk,n}\|^2}{\sigma^2 + \sum_{l \neq b} a_{il} P_l G_{lk} \|h_{lk,n}\|^2} \quad (2)$$

<sup>1</sup>Cell and BS are used interchangeable in this paper

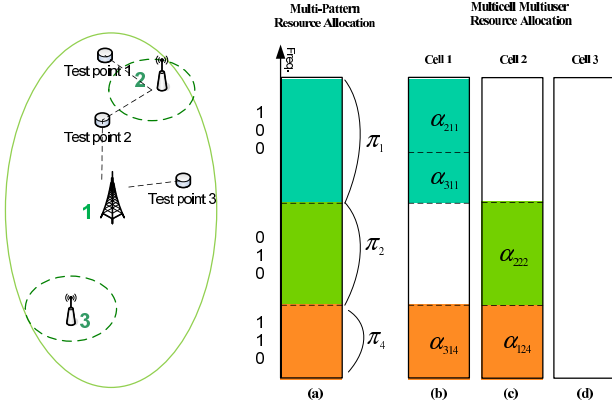


Fig. 1. An illustration of multi-pattern resource allocation in a HetNet consisting of one macro (Cell 1), two small cells (Cells 2 and 3), and 3 test points. Bar chart (a) gives a result of spectrum allocation among all 7 patterns, where  $\pi_i$  is the fraction allocated to pattern  $i$ . As shown,  $\pi_3 = \pi_5 = \pi_6 = \pi_7 = 0$ . Result of (a) can be directly translated into spectrum allocation among different cells as shown in (b), (c), and (d) respectively, where forbidden frequency resources are indicated in white color.  $\alpha_{kbi}$  denotes the fraction allocated to test point  $k$  by BS  $b$  under pattern  $i$ . In the shown result, Cell 1 serves test points 2 and 3, and Cell 2 serves points 1 and 2.

where  $a_{ib}$  is the cell activation indicator as given in (1),  $P_b$  is the PSD of BS  $b$ ,  $\sigma^2$  is the received noise PSD. We denote the channel gain between BS  $b$  and test point  $k$  over the  $n$ -th frequency resource by  $\sqrt{G_{bk}}h_{bk,n}$  where  $G_{bk}$  is the large-scale coefficient including antenna gain, path loss and shadowing, and  $h_{bk,n}$  accounts for the small-scale fading. We assume  $\{h_{bk,n}, \forall b, \forall k, \forall n\}$  are independent and identically distributed (i.i.d.). Hence, the ergodic rate of test point  $k$  served by the  $b$ -th BS under pattern  $i$  can be written as

$$\bar{r}_{kbi} = \alpha_{kbi} \underbrace{W \mathbb{E}_{\mathbf{h}} [\log_2 (1 + \text{SINR}_{kbi})]}_{\triangleq r_{kbi}} = \alpha_{kbi} r_{kbi} \quad (3)$$

where  $W$  is the system bandwidth,  $\alpha_{kbi}$  denotes the fraction of bandwidth that BS  $b$  allocates to test point  $k$  under pattern  $i$ ,  $\mathbf{h} \triangleq (h_{1k,n}, h_{2k,n}, \dots, h_{Bk,n})$ .

Finally, the total rate of test point  $k$  is obtained as

$$R_k = \sum_{i \in \mathcal{I}} \sum_{b \in \mathcal{B}} \bar{r}_{kbi} = \sum_{i \in \mathcal{I}} \sum_{b \in \mathcal{B}} \alpha_{kbi} r_{kbi}. \quad (4)$$

Note that  $r_{kbi}$  can be pre-calculated using (3) and hence treated as constants during the resource optimization. In (4), single-BS association restriction is not enforced, i.e., test point  $k$  is allowed to be connected to multiple BSs. The physical interpretation of this relaxation could be that the signals for test point  $k$  from multiple BSs are encoded and decoded separately by treating the signals from all BSs except one as interference (see (2)), and then the rate contributions from multiple BSs are summed up to give the final rate of test point  $k$ . We will show in Section V-B that this relaxation turns out to be really tight, in the sense that almost all of the test points are associated with single BS as the result of optimization.

### B. Energy consumption model

As mentioned previously, the BS usage is defined as  $\rho_b = \sum_k \sum_i \alpha_{kbi}$ ,  $\forall b$ . A typical power consumption model

for BSs consists of two types of power consumption: fixed power consumption and dynamic power consumption that is proportional to BS's utilization [2]. Denote by  $P_b^{\text{OP}}$  the maximum operational power of BS  $b$  if it is fully utilized (i.e.,  $\rho_b = 1$ ), which includes power consumption for transmit antennas as well as power amplifier, cooling equipment and so on. We can then express the total power consumption by all BSs as

$$P^{\text{tot}} = \sum_{b \in \mathcal{B}} [(1 - q_b) \rho_b P_b^{\text{OP}} + q_b |\rho_b|_0 P_b^{\text{OP}}] \quad (5)$$

where  $q_b \in (0, 1]$  is the portion of the fixed power consumption for BS  $b$  as long as it is switched on, and  $|x|_0$  is the function that takes the value of 0 if  $x = 0$  or the value 1 otherwise (i.e.,  $\ell_0$ -norm applied to a scalar). Note that by setting  $q_b = 1$  we arrive at a constant energy consumption model considered in [3], [14], which is a reasonable assumption for macro BSs. However, the small BSs such as pico or femto BSs may have smaller value of  $q_b$  because they do not usually have a big power amplifier or cooling equipment.

## III. RATE-CONSTRAINED ENERGY SAVING

### A. Problem formulation

The joint optimization of cell activation, user association and interference coordination via channel assignment for network energy saving can be formulated as

$$\begin{aligned} & \underset{\alpha, \pi}{* \text{ minimize}} \quad P^{\text{tot}} = \sum_{b \in \mathcal{B}} [(1 - q_b) \rho_b P_b^{\text{OP}} + q_b |\rho_b|_0 P_b^{\text{OP}}] \quad (6a) \end{aligned}$$

$$\text{subject to} \quad \rho_b = \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{I}} \alpha_{kbi}, \quad \forall b \quad (7)$$

$$\sum_{i \in \mathcal{I}} \sum_{b \in \mathcal{B}} \alpha_{kbi} r_{kbi} \geq d_k, \quad \forall k \quad (8)$$

$$\sum_{k \in \mathcal{K}} \alpha_{kbi} \leq \pi_i, \quad \forall b, \quad \forall i \quad (9)$$

$$\sum_{i \in \mathcal{I}} \pi_i = 1 \quad (10)$$

$$\pi_i \geq 0, \quad \forall i, \quad \alpha_{kbi} \geq 0, \quad \forall k, b, i \quad (11)$$

where (8) specifies the user demand of all test points, and all variables and parameters have been explained in Section II.

The difficulty of solving (6) lies in two facts. The first is the combinatorial objective function involving the  $\ell_0$ -norm. The other is that the number of all possible patterns in the network grows exponentially with the number of cells as  $2^B$ , resulting in huge problem dimension for a moderate-sized network. Fortunately, the following Proposition 1 identifies that only a small number of patterns out of  $2^B$  are needed for resource allocation to achieve the optimality.

**Proposition 1:** There exists an optimal solution to problem (6) that activates at most  $K + B + 1$  patterns, i.e.,  $|\{i \in \mathcal{I} : \pi_i > 0\}| \leq K + B + 1$ , where  $K$  and  $B$  are the number of users and number of cells in the network, respectively.

*Proof:* The proof is given in Appendix. ■

Proposition 1 indicates that if we know the set of most important patterns beforehand, the complexity of solving the resource allocation problem (6) can be significantly reduced

by restricting the candidate patterns to this set. Section IV-C suggests a practical guideline for pre-selecting candidate patterns. The effectiveness of this selection criterion in terms of energy saving will be evaluated in Section V.

### B. Feasibility test

Before describing the proposed method to solve (6), we introduce a rate balancing scheme to test the feasibility of (6) as follows. First, the rate requirement can be expressed using a normalized vector  $\beta = (\beta_1, \dots, \beta_K)^T$  where  $\beta_k = \frac{d_k}{\sum_{k \in \mathcal{K}} d_k}$ . Then the feasibility of problem (6) can be determined by solving the following rate balancing problem:

$$* \underset{\alpha, \pi, R_{\text{sum}}}{\text{minimize}} \quad -R_{\text{sum}} \quad (12a)$$

$$\text{subject to} \quad \beta_k R_{\text{sum}} - \sum_{i \in \mathcal{I}} \sum_{b \in \mathcal{B}} \alpha_{kbi} r_{kbi} \leq 0, \quad \forall k \quad (13)$$

$$(\alpha, \pi) \in \mathcal{X} \quad (14)$$

where  $\mathcal{X}$  is defined by (9), (10) and (11). Let  $R_{\text{sum}}^*$  denote the optimal value of problem (12). The original problem (6) is feasible if and only if  $R_{\text{sum}}^* \geq \sum_k d_k$ .

Note that problem (12) is a linear optimization problem and always feasible. It can be efficiently solved by, e.g., interior-point methods, using off-the-shelf solvers, if the problem dimension  $\mathcal{O}(IKB)$  is small. However, it is also desirable to solve (12) by involving a large number of patterns. This could happen when we consider all possible  $2^B$  patterns in order to calculate an optimal performance benchmark in a reasonable-sized network, or when the pre-selection still results in lots of candidate patterns for a large-scale network. In such case, the existing interior-point solvers, such as SeDuMi and SDPT3, cannot be applied, since they typically have cubic computational complexity in the problem dimension [15]. Fortunately, the problem has an interesting structure that facilitates a tailored cutting plane method to solve the dual problem.

The dual objective function can be written as

$$g(\lambda) = \inf_{(\alpha, \pi) \in \mathcal{X}} \left\{ R_{\text{sum}} \left( \sum_k \lambda_k - 1 \right) - \sum_{k, b, i} \alpha_{kbi} \frac{r_{kbi} \lambda_k}{\beta_k} \right\} \quad (15)$$

where  $\lambda = (\lambda_1, \dots, \lambda_K)^T$  with  $\lambda_k$  being the multiplier for the  $k$ -th inequality constraints in (13). This function is unbounded unless  $\sum_k \lambda_k = 1$ . Therefore, the corresponding dual problem can be stated as

$$\underset{\lambda \in \Lambda}{\text{maximize}} \quad \underset{(\alpha, \pi) \in \mathcal{X}}{\text{minimize}} \quad - \sum_{k, b, i} \alpha_{kbi} \frac{r_{kbi} \lambda_k}{\beta_k} \quad (16)$$

where  $\Lambda = \{\lambda : \sum_k \lambda_k = 1, \lambda_k \geq 0, \forall k\}$ . Since the primal problem (12) is feasible, strong duality holds for this linear program [16, Ch.5]. So (12) can be alternatively solved by the dual problem (16). The following Proposition 2 plays an important role for developing efficient solving algorithms.

**Proposition 2:** The inner minimization of problem (16) with fixed  $\lambda$  has a closed-form solution, which is

$$\bar{\alpha}_{kbi} = \begin{cases} 1 & \text{if } i = \bar{i}, k = \bar{k}(b, \bar{i}) \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

and

$$\bar{\pi}_i = \begin{cases} 1 & \text{if } i = \bar{i} \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

where

$$\bar{k}(b, i) = \arg \min_k \hat{r}_{kbi} \quad (19)$$

$$\bar{i} = \arg \min_i \sum_b \hat{r}_{\bar{k}(b, i)bi} \quad (20)$$

with

$$\hat{r}_{kbi} = -\frac{r_{kbi} \lambda_k}{\beta_k}. \quad (21)$$

*Proof:* The inner minimization of (16) with respect to  $(\alpha, \pi)$  can be rewritten as the following inner-outer formulation:

$$\underset{\pi_i \geq 0, \sum_i \pi_i = 1}{\text{minimize}} \quad \underset{\alpha_{kbi} \geq 0, \sum_k \alpha_{kbi} \leq \pi_i}{\text{minimize}} \quad \sum_k \sum_i \sum_b \alpha_{kbi} \hat{r}_{kbi} \quad (22)$$

where  $\hat{r}_{kbi}$  is defined in (21). Since  $\hat{r}_{kbi} \leq 0$ , it is clear that the inner problem of (22) with respect to  $\alpha$  is solved by each BS exclusively allocating maximum allowable resources to the single user who benefits the most for each pattern, i.e.,

$$\alpha_{kbi} = \begin{cases} \pi_i & \text{if } k = \bar{k}(b, i) \\ 0 & \text{otherwise} \end{cases} \quad (23)$$

where  $\bar{k}$  is expressed as (19). Substituting the solution of (23) back to (22), we arrive at the following problem:

$$\underset{\pi_i \geq 0, \sum_i \pi_i = 1}{\text{minimize}} \quad \sum_i \pi_i \sum_b \hat{r}_{\bar{k}(b, i)bi} \quad (24)$$

which is solved by pooling all resources to one pattern. So we obtain the solutions of (20) and (18), hence (17). ■

By means of introducing an auxiliary variable  $z$ , the dual problem (16) can be equivalently expressed as

$$* \underset{\lambda \in \Lambda, z}{\text{maximize}} \quad z \quad (25a)$$

$$\text{subject to} \quad - \sum_{k, b, i} \alpha_{kbi} \frac{r_{kbi} \lambda_k}{\beta_k} \geq z, \quad \forall (\alpha, \pi) \in \mathcal{X} \quad (26)$$

which unfortunately has infinitely many constraints in (26). The cutting plane algorithm solves an approximation at every iteration by considering only finite number of constraints and then refines this approximation by adding more constraints (cuts) for next iterations. Specifically, the following *master problem* is solved during the  $l$ -th iteration, given  $\alpha^{(0)}, \dots, \alpha^{(l-1)} \in \mathcal{X}$ :

$$\underset{\lambda \in \Lambda, z}{\text{maximize}} \quad z \\ \text{subject to} \quad - \sum_{k, b, i} \lambda_k \frac{\alpha_{kbi}^{(j)} r_{kbi}}{\beta_k} \geq z, \quad \forall j \in \{0, 1, \dots, l-1\}. \quad (27)$$

Let  $(\lambda^{(l)}, z^{(l)})$  be an optimal solution to the above problem (27). Then we have  $z^{(l)} \geq z^*$ , where  $z^*$  is the optimal value of the original dual problem (25), because the problem (27) has less restrictive constraints. In order to check whether  $(\lambda^{(l)}, z^{(l)})$  is also an optimal solution to the original dual

problem, we need to solve the inner minimization of problem (16) for given  $\lambda^{(l)}$ :

$$\underset{(\alpha, \pi) \in \mathcal{X}}{\text{minimize}} \quad - \sum_{k, b, i} \alpha_{kbi} \frac{r_{kbi} \lambda_k^{(l)}}{\beta_k}, \quad (28)$$

i.e., the dual function  $g$  is evaluated at  $\lambda^{(l)}$ . Let  $(\alpha^{(l)}, \pi^{(l)})$  be an optimal solution to the problem (28). If  $g(\lambda^{(l)}) \geq z^{(l)}$ , then  $(\lambda^{(l)}, z^{(l)})$  is an optimal solution to (25) (hence (16)). Otherwise,  $(\lambda^{(l)}, z^{(l)})$  is not a solution to the dual problem since it violates the constraint in (26) for  $\alpha = \alpha^{(l)}$ . In this case, the master problem of next iteration will be refined by adding a cut, i.e., adding  $\alpha^{(l)}$  to the current collection of points  $\alpha^{(0)}, \dots, \alpha^{(l-1)}$ .

Problems (27) and (28) are iteratively solved to find an optimal dual solution  $(\lambda^*, z^*)$ . The difficulty with huge dimension has now been encapsulated in problem (28) and nicely resolved thanks to the Proposition 2. The master problem (27) is a linear program with small dimension (not involving  $2^B$  term) that can be trivially solved using any standard solver. The whole algorithm is summarized in Algorithm I.

TABLE I  
ALGORITHM I: SOLVING RATE BALANCING PROBLEM (12) BY DUAL CUTTING PLANE

1: **Initialization:** Any point  $(\alpha, \pi) \in \mathcal{X}$  can be used as initial point. To obtain a sparse solution in particular, we choose  $(\alpha^{(0)}, \pi^{(0)})$  by activating a single pattern randomly and single-BS association for all test points. Set  $l = 0$ ;  
 2: **repeat**  
 3:    $l = l + 1$ ;  
 4:   Solve (27) by a standard primal-dual interior-point solver to obtain  $\lambda^{(l)}$  and  $z^{(l)}$ ;  
 5:   Solve (28) by Proposition 2 to obtain  $(\alpha^{(l)}, \pi^{(l)})$  and  $g(\lambda^{(l)})$ ;  
 6: **until**  $g(\lambda^{(l)}) \geq z^{(l)}$ ;  
 7: Reconstruct an optimal primal solution according to (29) and (30).

The complexity saving of the proposed algorithm in comparison to standard interior-point solvers can be briefly analyzed as follows. If problem (12) is directly solved by interior-point methods, the complexity is roughly  $\mathcal{O}(I^3 K^3 B^3)$ . By contrast, every iteration of the proposed algorithm requires finding a solution to (28) by Proposition 2, and a solution to (27) by interior-point solvers. Specifically, solving (28) requires  $\mathcal{O}(IKB)$ , while the complexity of solving (27) depends on the number of constraints in (27), which is increased by one inequality per iteration. Our numerical results suggest that the number of iterations is proportional to  $K$  (see footnote<sup>2</sup> for possible reasons). Consequently, it is safe to bound the complexity of solving (27) as  $\mathcal{O}(K^3)$  per iteration. Hence, the overall complexity of the proposed algorithm is  $\mathcal{O}(IK^2 B + K^4)$ , much lower than directly applying interior-point solvers to the original problem.

Finally, after the dual problem is solved by the proposed algorithm, the primal solution can be recovered as follows

<sup>2</sup>The reason that Algorithm I converges before the number of constraints in (27) grows significantly large is due to the inherent sparse structure of the solution. As identified by Proposition 1, the solution only activates a small number of patterns even if all possible patterns are candidate ones. Since the proposed algorithm activates one pattern per iteration (see (18)), the total number of iteration is unsurprisingly much lower than  $|I|$  if  $|I|$  is large.

[17, Ch.6]:

$$\alpha^* = \sum_{j=0}^{l-1} \kappa_j \alpha^{(j)} \quad (29)$$

$$\pi^* = \sum_{j=0}^{l-1} \kappa_j \pi^{(j)} \quad (30)$$

where  $\kappa_j$  with  $j = 0, \dots, l-1$  are the dual variables associated with the inequality constraints of (27), which are typically available as a by-product if we solve the problem (27) by a standard interior-point solver.

### C. Solving the energy saving problem with rate requirement

We now turn the attention to solving (6) if it is feasible. One popular approach to handle the  $\ell_0$ -norm term is the  $\ell_1$ -norm approximation. Applying this technique to (6), we obtain

$$* \underset{(\alpha, \pi) \in \mathcal{X}}{\text{minimize}} \quad \sum_{b \in \mathcal{B}} P_b^{\text{OP}} \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{I}} \alpha_{kbi} \quad (31a)$$

$$\text{subject to} \quad \sum_{i \in \mathcal{I}} \sum_{b \in \mathcal{B}} \alpha_{kbi} r_{kbi} \geq d_k, \quad \forall k \quad (32)$$

where  $\mathcal{X}$  is defined by (9), (10) and (11).

The solutions obtained from (31) can be further improved by applying so-called *reweighted  $\ell_1$ -norm minimization methods* [11], originally proposed to enhance the data acquisition in compressed sensing. It is known that for nonnegative scalar  $x \geq 0$ ,  $|x|_0 = \lim_{\epsilon \rightarrow 0} \frac{\log(1+x\epsilon^{-1})}{\log(1+\epsilon^{-1})}$  [3]. With a small design parameter  $\epsilon > 0$ , we neglect the limit and then approximate the  $\ell_0$ -norm as

$$|x|_0 \approx \frac{\log(1+x\epsilon^{-1})}{\log(1+\epsilon^{-1})}. \quad (33)$$

Relying on (33) and ignoring unnecessary constants, the problem (6) can be approximately solved by the following problem:

$$* \underset{(\alpha, \pi) \in \mathcal{X}}{\text{minimize}} \quad \sum_{b \in \mathcal{B}} \left[ (1 - q_b) P_b^{\text{OP}} \rho_b + \frac{q_b P_b^{\text{OP}} \log(\epsilon + \rho_b)}{\log(1 + \epsilon^{-1})} \right] \quad (34a)$$

$$\text{subject to} \quad \rho_b = \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{I}} \alpha_{kbi}, \quad \forall b \quad (35)$$

$$\sum_{i \in \mathcal{I}} \sum_{b \in \mathcal{B}} \alpha_{kbi} r_{kbi} \geq d_k, \quad \forall k. \quad (36)$$

Note that (34) is a continuous problem unlike the one in (6) involving combinatorial terms. However, (34) is not a convex problem since it minimizes a concave function. Fortunately, it falls into the framework of difference-of-convex (DC) functions and therefore can be efficiently solved by the convex-concave procedure [18].

Specifically, by applying the first-order Taylor expansion to the objective function in (34) at the point  $\rho^{(t-1)}$  obtained in  $(t-1)$ -th iteration, we arrive at the following problem for the  $t$ -th iteration:

$$* \underset{(\alpha, \pi) \in \mathcal{X}}{\text{minimize}} \quad \sum_{b \in \mathcal{B}} w_b^{(t)} \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{I}} \alpha_{kbi} \quad (37a)$$

$$\text{subject to} \quad \sum_{i \in \mathcal{I}} \sum_{b \in \mathcal{B}} \alpha_{kbi} r_{kbi} \geq d_k, \quad \forall k \quad (38)$$

where the weight

$$w_b^{(t)} = (1 - q_b)P_b^{\text{OP}} + \frac{q_b P_b^{\text{OP}}}{\log(1 + \epsilon^{-1})(\epsilon + \rho_b^{(t-1)})} \quad (39)$$

with

$$\rho_b^{(t-1)} = \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{I}} \alpha_{kbi}^{(t-1)}. \quad (40)$$

The convergence can be characterized as follows:

**Proposition 3:** Any limiting point of  $(\alpha^{(t)}, \pi^{(t)})$  generated by the above convex-concave procedure as  $t \rightarrow \infty$  is a stationary point of problem (34).

*Proof:* We first eliminate the equality constraint (35) by substituting it into the objective (34), and then denote the rest of constraint set  $(\mathcal{X}$  with (36)) by  $\mathcal{Y}$ . It can be easily verified that  $\mathcal{Y}$  is compact (closed and bounded). According to Remark 7 in [29], our problem satisfies all conditions of Theorem 4 in [29]. By applying this theorem, the proposition is proved. ■

In practice, the reweighted  $\ell_1$  method converges typically within 6-10 iterations to a desirable accuracy and the largest improvement in sparsity is obtained in the first few iterations.

Problem (37) is a linear program. Like in Section III-B, it can be tackled in the dual domain by the cutting plane method, resulting in similar complexity saving as explained in Section III-B. In Section V-B, we will compare the running time of the proposed algorithm to that of a commercial solver by simulation.

By dualizing the constraint of (38), we can express the dual function as

$$h(\mu) = \inf_{(\alpha, \pi) \in \mathcal{X}} \left\{ \sum_{k,b,i} \alpha_{kbi} w_b^{(t)} - \sum_{k,b,i} \alpha_{kbi} r_{kbi} \mu_k + \sum_k d_k \mu_k \right\} \quad (41)$$

where  $\mu = (\mu_1, \dots, \mu_K)^T$  is the Lagrangian multiplier.

Following the idea presented in Section III-B, we formulate the master problem as

$$\begin{aligned} & \underset{\mu \geq 0, z}{\text{maximize}} \quad z \\ & \text{subject to} \quad \sum_{k,b,i} \alpha_{kbi}^{(t,j)} w_b^{(t)} - \sum_{k,b,i} \alpha_{kbi}^{(t,j)} r_{kbi} \mu_k + \sum_k d_k \mu_k \geq z, \\ & \quad \quad \quad \forall j \in \{0, \dots, l-1\} \end{aligned} \quad (42a) \quad (42b)$$

and the inner problem as

$$\underset{(\alpha, \pi) \in \mathcal{X}}{\text{minimize}} \quad \sum_{k,b,i} \alpha_{kbi} \left( w_b^{(t)} - r_{kbi} \mu_k^{(l)} \right) + \sum_k d_k \mu_k^{(l)} \quad (43)$$

respectively, where we denote the solution to (42) by  $(\mu^{(l)}, z^{(l)})$  and the solution to (43) by  $(\alpha^{(t,l)}, \pi^{(t,l)})$ . The master problem (42) is refined for the next iteration by adding  $\alpha^{(t,l)}$  to the constraint (42b). In this way, we iteratively solve (42) and (43) until  $h(\mu^{(l)}) \geq z^{(l)}$ , implying that we have solved the problem (37) in the dual domain. Then we can find the primal solution following the same idea of (29) and (30) as:

$$\alpha^{(t)} = \sum_{j=0}^{l-1} \kappa_j \alpha^{(t,j)} \quad (44)$$

$$\pi^{(t)} = \sum_{j=0}^{l-1} \kappa_j \pi^{(t,j)} \quad (45)$$

where  $\kappa_j$  with  $j = 0, \dots, l-1$  are the dual variables corresponding to the inequality constraints of (42b), which are available if we solve the master problem (42) by off-the-shelf interior-point solvers.

Finally, the outermost iteration is to adjust the weights according to (39) and (40) and then the problem (37) is solved again with the new weights. We summarize the proposed approach in Algorithm II.

TABLE II  
ALGORITHM II: ENERGY SAVING WITH THE USER RATE CONSTRAINT

1: <b>Feasibility check:</b> Solve the rate balancing problem (12) by the methods described in Section III-B to obtain $\alpha_{\text{bln}}^*, \pi_{\text{bln}}^*$ , and $R_{\text{sum}}^*$ . If $R_{\text{sum}}^* < \sum_k d_k$ , then the user rate constraint is infeasible; Else if $R_{\text{sum}}^* = \sum_k d_k$ , problem is solved; Otherwise we proceed to the next step;
2: <b>Initialization:</b> Outer iteration counter $t = 0$ , $\alpha^{(0)} = \alpha_{\text{bln}}^*$ , $\pi^{(0)} = \pi_{\text{bln}}^*$ ;
3: <b>repeat</b>
4: $t = t + 1$ ;
5:   Update the weights: If $t = 1$ , then $\{w_b^{(1)} = P_b^{\text{OP}}, \forall b\}$ as given in (31); otherwise calculate $\{w_b^{(t)}, \forall b\}$ according to (39) and (40);
6:   Initialize inner iteration counter $l = 0$ , $\alpha^{(t,0)} = \alpha_{\text{bln}}^*$ , $\pi^{(t,0)} = \pi_{\text{bln}}^*$ ;
7: <b>repeat</b>
8: $l = l + 1$ ;
9:     Solve (42) by a standard primal-dual interior-point solver to obtain $\mu^{(l)}, z^{(l)}$ ;
10:     Solve (43) by Proposition 4 to obtain $\alpha^{(t,l)}, \pi^{(t,l)}$ and $h(\mu^{(l)})$ ;
11: <b>until</b> $h(\mu^{(l)}) \geq z^{(l)}$ ;
12:     Reconstruct an optimal primal solution $(\alpha^{(t)}, \pi^{(t)})$ from $(\alpha^{(t,j)}, \pi^{(t,j)})$ , $j = 0, \dots, l-1$ according to (44) and (45);
13: <b>until</b> Objective (34) converges or the maximum number of iterations is reached.

The key enabler in Algorithm II is the following Proposition 4, which shares the similar spirit to Proposition 2:

**Proposition 4:** The problem (43) has a closed-form solution that can be expressed as

$$\alpha_{kbi}^{(t,l)} = \begin{cases} 1 & \text{if } i = \bar{i}, k = \bar{k}(b, \bar{i}), \text{ and } \tilde{r}_{kbi} < 0 \\ 0 & \text{otherwise} \end{cases} \quad (46)$$

and

$$\pi_i^{(t,l)} = \begin{cases} 1 & \text{if } i = \bar{i} \\ 0 & \text{otherwise} \end{cases} \quad (47)$$

where  $\bar{k}(b, i) = \arg \min_k \tilde{r}_{kbi}$  with  $\tilde{r}_{kbi} = w_b^{(t)} - r_{kbi} \mu_k^{(l)}$ , and  $\bar{i} = \arg \min_i \sum_b [\tilde{r}_{\bar{k}(b,i)bi}]^-$ , where  $[x]^- = \min(0, x)$ .

*Proof:* The proof can be obtained following the similar steps that we use to prove Proposition 2. The only difference is that  $\tilde{r}_{kbi}$  can now take positive values. If  $\min_k \tilde{r}_{kbi} > 0$  for the given  $b$  and  $i$ , then

$$\underset{\alpha_{kbi} \geq 0, \sum_k \alpha_{kbi} \leq \pi_i}{\text{minimize}} \quad \sum_{k,b,i} \alpha_{kbi} \tilde{r}_{kbi} \quad (48)$$

will result in  $\alpha_{kbi} = 0, \forall k$  for the given  $b$  and  $i$ . Moreover, if  $(\min_k \tilde{r}_{kbi}) = 0$  for the given  $b$  and  $i$ , we can also set  $\alpha_{kbi} = 0$  without affecting the optimality. Taking these two facts into account, we modify  $\alpha_{kbi}^{(t,l)}$  in (46) accordingly, and  $\bar{i}$  as well. ■

Several remarks to the Algorithm II are as follows.

**Remark 1:** The cutting plane method should be initialized with a strictly primal feasible solution, otherwise the master problem will become unbounded in the first iteration or

not work properly. The proposed Algorithm II checks the feasibility of the primal problem in step 1. If the prescribed rate constraints are feasible, the checking procedure guarantees to find a strictly primal feasible solution, which is used to initialize the cutting plane iteration (see step 6).

*Remark 2:* In step 5, this particular choice of the initial weighting matrix means that we solve the (unweighted)  $\ell_1$ -norm approximation problem (31) directly in the first iteration.

#### IV. UNIFIED STUDY OF BASELINE STRATEGIES

In this section, we show how to develop a unified view on a wide range of previous strategies. With this view, we can analyze and compare various resource management strategies in a unified framework.

##### A. Cell activation and user association without interference coordination

The cell activation and user association has been studied in [3], [4], where worst-case estimates of the user rates resulted from no intercell interference coordination are used to calculate the QoS requirements. The strategy presented in [3], [4] can be easily analyzed using the proposed framework in this paper. It corresponds to restricting the candidate pattern set to exactly one pattern: All-ON pattern (i.e., reuse-1 pattern). To compute the link rate under the reuse-1 pattern, we set  $\{a_{il} = 1, \forall l \in \mathcal{B}\}$  in (2) and then calculate the rate according to (3). Since there is only one allowed pattern for resource allocation, we can drop both the subscript  $i$  and the pattern allocation variable  $\pi$ , focusing on the two-dimension resource allocation variable  $\alpha_{kb}$  only.

Consequently, the user rate constrained energy saving problem of (6) boils down to

$$* \underset{\{\alpha_{kb}\}}{\text{minimize}} \quad P^{\text{tot}} = \sum_{b \in \mathcal{B}} [(1 - q_b)\rho_b P_b^{\text{OP}} + q_b|\rho_b|_0 P_b^{\text{OP}}] \quad (49a)$$

$$\text{subject to} \quad \rho_b = \sum_{k \in \mathcal{K}} \alpha_{kb}, \forall b \quad (50)$$

$$\sum_{b \in \mathcal{B}} \alpha_{kb} r_{kb} \geq d_k, \forall k \quad (51)$$

$$\sum_{k \in \mathcal{K}} \alpha_{kb} \leq 1, \forall b \quad (52)$$

$$\alpha_{kb} \geq 0, \forall k, b \quad (53)$$

which can be solved by the convex-concave procedure as described before. We will compare this strategy with our proposal in Section V.

##### B. Range expansion user association

Reference signal received power (RSRP) is adopted in LTE/LTE-A standards as a signal quality indicator [20]. In LTE-A HetNets, range expansion (RE) has been further introduced as a simple scheme to control the load distribution among pico and macro layers [21]. The basic mechanism is to add a positive bias (in dB) to the RSRP received from small cells when deciding the association of user equipments (UEs).

In the proposed framework, to express the association of the test point  $k$  according the RE rule, a binary association indicator  $s_{kb}$  can be introduced:

$$s_{kb} = \begin{cases} 1 & \text{if } b = \arg \max_{j \in \mathcal{B}} (\text{RSRP}_{jk} + \eta_j) \\ 0 & \text{otherwise} \end{cases}, \quad \forall k \quad (54)$$

where  $\text{RSRP}_{jk}$  is the received RSRP (in dBm) at the test point  $k$  from cell  $j$ , and  $\eta_j$  is the bias value (in dB) of cell  $j$ .

1) *Cell activation and interference coordination with fixed RE association*: In order to separate the impact of the RE user association rule from that of resource allocation, we can pre-define a set of RE biases and fix the user association according to (54). Then the cell activation and resource allocation problem with fixed user association can be formulated as

$$* \underset{\alpha, \pi}{\text{minimize}} \quad P^{\text{tot}} = \sum_{b \in \mathcal{B}} [(1 - q_b)\rho_b P_b^{\text{OP}} + q_b|\rho_b|_0 P_b^{\text{OP}}] \quad (55a)$$

$$\text{subject to} \quad \rho_b = \sum_{k \in \mathcal{K}_b} \sum_{i \in \mathcal{I}} \alpha_{kbi}, \quad \forall b \quad (56)$$

$$\sum_{i \in \mathcal{I}} \sum_{b \in \mathcal{B}} \alpha_{kbi} r_{kbi}^{\text{RE}} \geq d_k, \quad \forall k \quad (57)$$

$$\sum_{k \in \mathcal{K}_b} \alpha_{kbi} \leq \pi_i, \quad \forall b, \forall i \quad (58)$$

$$\sum_{i \in \mathcal{I}} \pi_i = 1 \quad (59)$$

$$\pi_i \geq 0, \quad \forall i, \quad \alpha_{kbi} \geq 0, \quad \forall k, b, i \quad (60)$$

where  $r_{kbi}^{\text{RE}} \triangleq s_{kb} r_{kbi}$  is the effective rate obtained by forcing the elements of  $r_{kbi}$  that are not allowed to associate due to the RE rule to zero, and  $\mathcal{K}_b = \{k \in \mathcal{K} : s_{kb} = 1\}$ .

The above formulation is almost the same as the problem (6) with the only two differences. First, the rate  $r_{kbi}$  in problem (6) is replaced by the effective rate  $r_{kbi}^{\text{RE}}$ . Second, the summations in (56) and (58) are over  $\mathcal{K}_b$  instead of  $\mathcal{K}$ . Interestingly, the following Proposition 5 justifies that  $\mathcal{K}_b$  can be replaced by  $\mathcal{K}$  without loss of optimality, meaning that the algorithms developed in Section III can be directly applied to solve (55).

*Proposition 5:* The problem (55) can be equivalently solved by replacing all  $\mathcal{K}_b$  with  $\mathcal{K}$ .

*Proof:* By replacing  $\mathcal{K}_b$  with  $\mathcal{K}$  in problem (55), we restrict the feasible set since  $\mathcal{K}_b \subseteq \mathcal{K}$ . However, doing so will not compromise the optimality. The reason is that BS  $b$  does not contribute any rate for users outside  $\mathcal{K}_b$  (due to the definition of  $r_{kbi}^{\text{RE}}$ ). Hence, we can set  $\alpha_{kbi} = 0$ , if  $k \notin \mathcal{K}_b$ . Formally, the proof is given as follows.

In this proof, we refer to the new problem where  $k \in \mathcal{K}_b$  in (56) and (58) has been replaced with  $k \in \mathcal{K}$  as the *reformulated problem*. Let  $(\alpha^{\text{new}}, \pi^{\text{new}})$  be the solution to this reformulated problem. Accordingly in the reformulated problem, define  $\rho_b^{\text{new}} = \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{I}} \alpha_{kbi}^{\text{new}}, \forall b \in \mathcal{B}$ . Note that  $(\alpha^{\text{new}}, \pi^{\text{new}})$  satisfies the original constraints from (57) to (60), since  $r_{kbi}^{\text{RE}} \geq 0, \mathcal{K}_b \subseteq \mathcal{K}$  and  $\alpha_{kbi} \geq 0$ . Next we prove that  $(\alpha^{\text{new}}, \pi^{\text{new}})$  must be the solution to problem (55) by contradiction.

Suppose this is not true, meaning that we can find another feasible point  $(\alpha^{\text{old}}, \pi^{\text{old}})$  in problem (55) such that  $P^{\text{tot}}(\rho^{\text{old}}) < P^{\text{tot}}(\rho^{\text{trim}})$ , where  $\rho^{\text{old}} = (\rho_1^{\text{old}}, \dots, \rho_B^{\text{old}})^T$  with



$\rho_b^{\text{old}} = \sum_{k \in \mathcal{K}_b} \sum_{i \in \mathcal{I}} \alpha_{kbi}^{\text{old}}$ , and  $\rho^{\text{trim}} = (\rho_1^{\text{trim}}, \dots, \rho_B^{\text{trim}})^T$  with  $\rho_b^{\text{trim}} = \sum_{k \in \mathcal{K}_b} \sum_{i \in \mathcal{I}} \alpha_{kbi}^{\text{new}}$ .

In such case, we can construct another point  $(\alpha', \pi')$  by choosing  $\alpha'_{kbi} = \begin{cases} 0 & \text{if } k \notin \mathcal{K}_b \\ \alpha_{kbi}^{\text{old}} & \text{otherwise} \end{cases}$  and  $\pi' = \pi^{\text{old}}$ , respectively. It can be easily seen that  $(\alpha', \pi')$  is also feasible in the reformulated problem. Define  $\rho'_b = \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{I}} \alpha'_{kbi}$ ,  $\forall b$ . We arrive at  $P^{\text{tot}}(\rho') = P^{\text{tot}}(\rho^{\text{old}}) < P^{\text{tot}}(\rho^{\text{trim}}) \leq P^{\text{tot}}(\rho^{\text{new}})$ . In other words, we find a feasible point  $(\alpha', \pi')$  in the reformulated problem that gives lower value of the objective function than  $(\alpha^{\text{new}}, \pi^{\text{new}})$ , which is contradictory to the optimality of  $(\alpha^{\text{new}}, \pi^{\text{new}})$  in the reformulated problem. ■

2) *Mapping the jointly optimized association to cell-specific biases*: In Section III, we propose and solve the coupled problem of optimizing the user association, cell activation and resource allocation. It is interesting to see how this jointly optimized user association can be mapped to the cell-specific biases. In other words, we would like to choose values of  $\eta_j, \forall j \in \mathcal{B}$  such that the user association based on the rule given by (54) leads to the association decisions derived in Section III by a joint optimization.

To achieve this goal, we propose to minimize a weighted mean square error of the association. Specifically, we aim to solve the following optimization problem:

$$\eta^* = \arg \min_{\eta \geq 0} \sum_{b \in \mathcal{B}} \omega_b \sum_{k \in \mathcal{K}} (s_{kb}(\eta) - s_{kb}^*)^2 \quad (61)$$

where  $\eta = (\eta_1, \dots, \eta_B)^T$ , the weight  $\omega_b$  is used to emphasize the different impacts of the association error on different cells (e.g., macro cells can have larger weights to account for the larger energy consumption if macro cells are forced to switch on due to errors in the user association),  $s_{kb}$  is related to  $\eta$  according to (54), the reference association  $s_{kb}^*$  is derived from the joint optimization given in Section III as  $s_{kb}^* = |\sum_{i \in \mathcal{I}} \alpha_{kbi}^*|_0, \forall k, \forall b$ , and  $\alpha_{kbi}^*$  is the solution to problem (6) using the Algorithm II.

Problem (61) is solved by a coordinate descent method in this paper, i.e., one-dimensional search for one bias value is performed at a time while keeping the rest of biases fixed. The iterative procedure is terminated if the objective cannot be further reduced. The objective function is guaranteed to converge, although the resulting solution is not necessarily globally optimal. The solution depends on the order of BSs being updated. Hence, we solve (61) several times with different updating orders, and then select the best one. Also note that the RE rule typically results in the single-BS association for each test points, but the solution derived by joint optimization in Section III may yield multiple-BS association for some test points. Hence, in general the minimum error in (61) is greater than zero. Nevertheless, as will be shown in Section V-D, the proposed method for solving (61) gives a set of nearly optimal cell-specific biases.

### C. Cell clustering for activation

In order to reduce the number of candidate patterns for resource allocation, we can group multiple cells into clusters. Then each cluster is regarded as one giant cell when

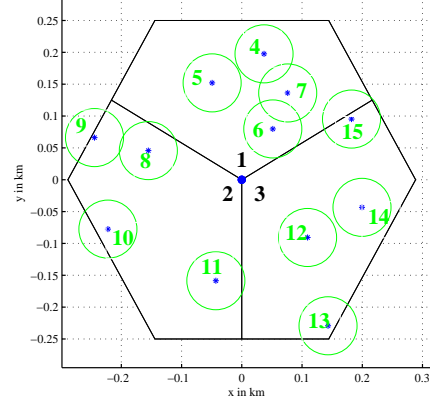


Fig. 2. A heterogeneous network consisting of 15 cells.

formulating interference patterns. In other words, all cells within the same cluster can only be simultaneously activated or deactivated.

The principle of forming a cluster is to group cells that do not interfere with each other or have very weak mutual interference into one cluster. In this way, simultaneously activating them will not significantly increase the network interference. In HetNets, pico BSs have low transmit power and antenna gain. So the interference among pico cells is expected to be low if they are deployed with reasonable inter-site distances. Therefore, we can group pico cells within one macro cell into one cluster. On the other hand, we separate the dominant mutually interfering cells (e.g., pico cells and its umbrella macro cell) into different clusters, such that the inter-cluster interference can be handled by resource allocation among clusters.

The framework proposed in Section III can be used to evaluate various cell clustering strategies from the energy saving perspective. Cell clustering results in a restricted set of candidate patterns. The user association and resource allocation can then be performed over this pattern set. The algorithms developed in Section III can be directly applied.

## V. PERFORMANCE EVALUATION

### A. Simulation setup

A network consisting 15 cells have been used in the simulations. Among these cells, there are 3 macro cells, each of which contains 4 randomly dropped pico cells, as shown in Fig. 2. The cells are labeled as

$$\underbrace{1, 2, 3}_{\text{macro cells}}, \underbrace{4, 5, 6, 7}_{\text{picos in cell 1}}, \underbrace{8, 9, 10, 11}_{\text{picos in cell 2}}, \underbrace{12, 13, 14, 15}_{\text{picos in cell 3}}.$$

The parameters for propagation modelling follow the suggestions in 3GPP evaluation methodology [22], and are summarized in Table III together with other system parameters. Based on the linear relationship between transmit power and operational power consumption<sup>3</sup>, we calculate the maximum

<sup>3</sup>We adopted the linear model in [23]:  $P_b^{\text{OP}} = \alpha_b P_b + \beta_b$ , where  $P_b$  is the transmit power for BS  $b$ ,  $\alpha_b = \frac{22.6}{3}$  and  $\beta_b = \frac{412.4}{3}$  W if  $b$  is a macro; otherwise  $\alpha_b = 5.5$  and  $\beta_b = 32$  W if  $b$  is a pico.



TABLE III  
NETWORK PARAMETERS.

Parameter	Description
bandwidth	10 MHz
Macro total Tx power	46 dBm
Macro $P^{OP}$ and $q_b$	439 W, 1
Pico total Tx power	30 dBm
Pico $P^{OP}$ and $q_b$	38 W, 0.5
Macro antenna gain	15 dB
Pico antenna gain	5 dB
Macro path loss	$128.1 + 37.6 \log_{10}(R)$
Pico path loss	$140.7 + 36.7 \log_{10}(R)$
Penetration loss	20 dB
Shadowing std. dev.	8dB(macro), 10dB(pico)
Shadowing corr. distance	25 m
Macrocell shadowing corr.	1 between cells
Picocell shadowing corr.	0.5 between cells
Fading model	No fast fading
Min. macro(pico)-UE dist.	35 m (10 m)
Min. macro(pico)-pico dist.	75 m (40 m)
Noise density and noise figure	-174 dBm/Hz, 9dB

TABLE IV  
NUMBER OF ACTIVE PATTERNS AFTER ALGORITHM II CONVERGES.

Rate requirement (Mbit/s)	0.1	0.5	1.0	1.5	2.0
50 test points	5	9	18	18	23
150 test points	6	25	43	47	infeasible

operational power  $P^{OP}$  as 439W and 38W for macro and pico BSs, respectively. We further assume each macro BS has a constant power consumption, i.e.,  $q_b = 1, \forall b \in \mathcal{B}_{\text{macro}}$ , and the fixed power consumption of a pico takes 50% of the maximum operational power, i.e.  $q_b = 0.5, \forall b \in \mathcal{B}_{\text{pico}}$ . Note that these assumptions are made for providing concrete numerical results, and they are not from the restriction of our formulation. For reweighted  $\ell_1$ -norm minimization, we set  $\epsilon = 10^{-6}$  and the maximum number of iteration is 10.

### B. Performance of the proposed algorithm

Fig. 3 and Fig. 4 report the network power consumption and number of active BSs, respectively, obtained by the proposed Algorithm II. The results are plotted versus the rate requirement of the test points, where 50 and 150 test points are uniformly distributed within the network after dropping the picos, and all test points are assumed to have the same rate requirement for simplicity.

As shown in Figs. 3 and 4, the network power consumption, as well as the number of active BSs, increases with the user rate requirement. The maximum feasible rate requirement for 50 test points in the considered network realization is  $d_k = 4.32\text{Mbit/s}$ ,  $\forall k$ , while it reduces to 1.58Mbit/s if 150 test points need to be supported. In Figs. 3 and 4, the rate requirements are chosen uniformly between 0.01Mbit/s and the maximum feasible rates. Interestingly, we observe 3 steep increases in the power consumption in Fig. 3 for both 50-test-point-case and 150-test-point-case. Actually, each of these jumps corresponds to activating one macro BS. Note that the proposed algorithm successfully deactivates the macro BSs for power saving when the rate requirement is small or moderate. It activates macro BSs only if necessary.

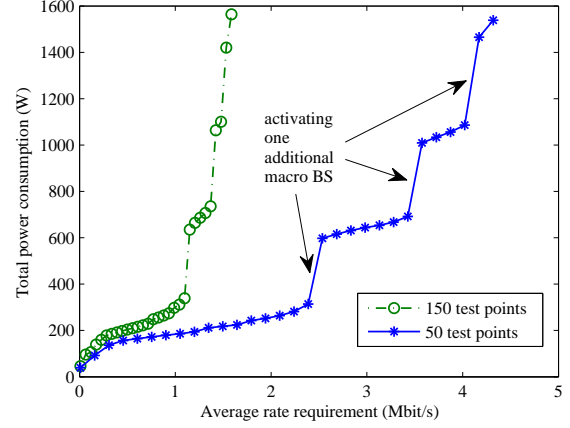


Fig. 3. Network power consumption achieved by the proposed Algorithm II, where all test points are assumed to have the same rate requirement.

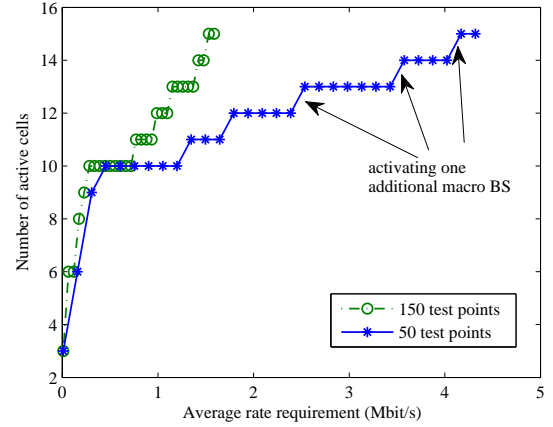


Fig. 4. Number of active cells achieved by the proposed Algorithm II, where all test points are assumed to have the same rate requirement.

To verify the sparsity structure of the solution indicated by Proposition 1, we show in Table IV the number of active patterns after Algorithm II converges when all possible  $2^{15}$  patterns are considered as the candidates. As shown, the solution found by the proposed algorithm indeed allocates resources to a small number of patterns. Most of the candidate patterns have not been used. The number of active patterns slightly increases with the rate requirement, but is less than the bound established in Proposition 1.

We further show the average number of serving BSs for each test point in Table V. As shown, although problem formulation of (6) does not enforce single-BS association constraints, almost all test points are associated with single BS as the result of optimization. In the simulated cases, multiple-associated test points only occur in three scenarios: 50 test points with demands of 2.0Mbit/s, 150 test points with 1.0Mbit/s and 1.5Mbit/s. In these three scenarios, only 1 out of 50, 4 out of 150 and 3 out of 150 are multiple-associated test points, respectively. The number of serving BSs for these multiple-associated test points are two.

To enforce single association for ALL test points, we can introduce the binary association indicator  $s_{kb}$  and apply range

TABLE V  
AVERAGE NUMBER OF SERVING BSS PER TEST POINT.

Rate requirement (Mbit/s)	0.1	0.5	1.0	1.5	2.0
50 test points	1	1	1	1	1.020
150 test points	1	1	1.027	1.020	infeasible

TABLE VI  
ALGORITHM RUNNING TIME.

Number of candidate patterns	19	$2^6$	$2^9$	$2^{15}$
Proposed algorithm (sec)	4.2	10.2	13.8	313
Gurobi solver (sec)	0.3	1.2	19.6	6346

expansion according to (54), where the cell-specific biases can be obtained from the optimal user resource parameter  $\alpha_{kbi}$ , as given by (61). This range expansion based single association will facilitate the implementation in practice. The performance will be evaluated in the following sections.

Finally, we compare the running time of Algorithm II to that of replacing steps 6 to 12 with state-of-the-art commercial solver, Gurobi [30] (with barrier method selected), and report the results in Table VI. In the simulated case, 50 test points are uniformly distributed in the network with the same demand of 200kbit/s. The results are averaged over 10 random drops. The algorithms are executed in Matlab 2014 on an Intel Core i7 2.2GHz quad-core computer with 8GB RAM. We apply clustering to obtain different number of candidate patterns as shown in Table VI (see Section V-C for details of feature pattern selection). Compared to the industrial-strength solver, our algorithm with a unsophisticated implementation starts to achieve some gains as the problem dimension grows. In particular, when  $2^{15}$  patterns are considered, a significant saving in running time is observed. To conclude, the proposed algorithm provides a feasible way to calculate the benchmark considering all  $2^B$  patterns in a reasonable-sized network. It can also be applied to larger networks over a set of preselected patterns by clustering, still achieving complexity saving.

### C. Comparing different strategies

In this subsection, we illustrate how to use the proposed framework to compare various existing user association and resource allocation strategies in terms of network power consumption.

The first strategy is the proposed algorithm in this paper, where all  $2^{15}$  patterns are considered in the candidate pattern set.

The second is the *Reuse-1* scheme [3], [4], which can be cast into the proposed framework by restricting the candidate pattern to a single reuse-1 pattern (see Section IV-A).

The third one is the *Pre-selected feature patterns* scheme. The idea is to group pico BSs within one macro cell into one cluster when formulating interference patterns (see discussions in Section IV-C). In such way, the number of possible patterns has been reduced from  $2^{15}$  to  $2^6$ . We further restrict the candidate patterns to the following *four* by switching on:  $\{p1, p2, p3\}$ ,  $\{1, p2, p3\}$ ,  $\{2, p1, p3\}$ , and  $\{3, p1, p2\}$ , where  $p1$ ,  $p2$ ,  $p3$  denote the pico clusters within cells 1, 2, and 3, respectively. This was a suggested feature pattern set in [9] achieving

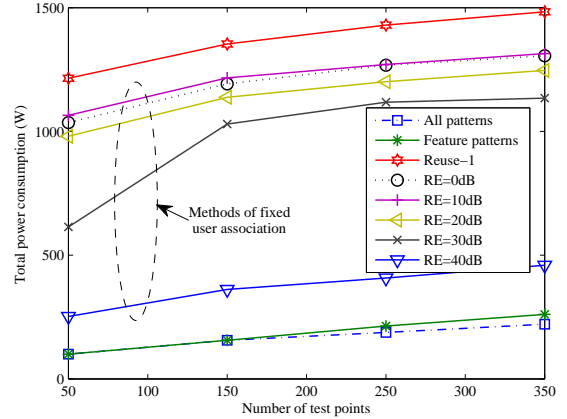


Fig. 5. Network power consumption of different schemes. All test points have the same rate requirement  $d_k = 200\text{kbit/s}$ , and uniformly distributed in the network. The results are averaged over ten realizations for each given number of test points.

nearly optimal rate utility maximization. In addition to these four patterns, we add 15 patterns, each one activating one single cell, in order to increase the granularity of interference characterization.

The fourth strategy is to separate the user association from the joint optimization and apply the simple range expansion rule. In the evaluation, we set the macro bias to zero, and the same bias for all the pico BSs, choosing from 0, 10, 20, 30 and 40 dB.<sup>4</sup> All possible patterns are included in the resource allocation.

The results are given in Fig. 5. As shown, the proposed algorithm considering all patterns achieves the minimum power consumption for all the given test-point-cases. We can use it as a benchmark for quantifying other strategies where the patterns are somehow restricted or resource optimization is decoupled from the user association (by RE rules).

The feature-pattern scheme achieves almost the same power saving performance as using all patterns. This is because there is little loss in the user achievable rates by characterizing interference using this feature pattern set, as shown in [9]. When it is used for activating cells to satisfy the user rate demand, it is not surprising to achieve close-to-benchmark performance.

The existing strategy based on reuse-1, on the other hand, achieves the worst power saving performance among all the methods in comparison. This is because the interference coupling has not been taken into account when (de)activating cells. It completely neglects the fact that muting some BSs can reduce the interference and hence increase the user rate. Therefore, the resulting BS activation and user association decisions are highly sub-optimal. For example, as shown in Fig. 5, the proposed algorithms (all patterns and feature patterns) only

<sup>4</sup>In the current LTE networks, the maximum bias for pico range expansion is typically restricted to 15 dB. Too aggressive bias will potentially cause the control channel failure. In this paper, we do not consider this restriction when investigating the full potential of the range expansion scheme. Our consideration can be justified by assuming a split between U-plane and C-plane, and there exists some control channel protection mechanism, e.g. control channels of pico and macro are deployed on the orthogonal resources.

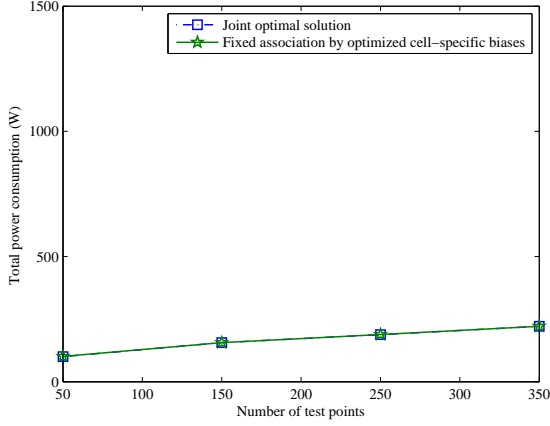


Fig. 6. Comparison of the jointly optimized solution with the fixed association using optimized cell-specific bias. All test points have the same rate requirement  $d_k = 200\text{ kbit/s}$ , and uniformly distributed in the network. The results are averaged over ten realizations for each given number of test points.

need 12% of the power consumption of the reuse-1 scheme to support 150 test points with 200kbit/s.

We can also observe from Fig. 5 that the methods with the fixed RE association achieve the performance between the benchmark and the worst reuse-1 method. Moreover, by increasing the pico biases from 0 dB to 40 dB, more and more test points are offloaded onto the pico cells, increasing the opportunity of deactivating macro cells. Hence, it reduces the power consumption. However, we still observe a considerable performance loss even a large bias of 40 dB is used for pico cells, compared to the benchmark. This is because all pico BSs have to use the same bias value. We will make this point clearer by computing cell-specific biases in the next subsection.

#### D. Cell-specific bias

In Table VII, we show the cell-specific bias obtained from the jointly optimized solution for a typical network realization using the algorithm described in Section IV-B2. As can be seen, the minimization of network power consumption requires different bias for different cells. This is in sharp contrast to the conclusions derived in [13] from the perspective of load balancing, where the same bias per-tier resulted in almost no performance loss in the network rate utility. This is because biasing for energy saving is targeted at a different goal from load balancing. In order to deactivate some BSs, biasing here is used as a mechanism to concentrate users to a small set of cells.

By using the derived cell-specific bias from the jointly optimized solution, we again solve the network power minimization problem of (55) with fixed user association, and report the results in Fig. 6. As shown, the previous gap in Fig. 5 between the benchmark and the scheme with fixed RE association is now closed.

#### E. Modeling aspects and practical implementation of the solution

During the relatively long decision period considered in this paper, users could join the network and then leave after being served. Hence, we adopt test points as an abstract concept to represent demands of users in a given region, as in [4]. The test points can be chosen from typical user locations, or we can simply partition the geographic region into pixels, within each pixel radio propagation being considered uniform and then each pixel becomes one test point.

In the proposed model, the demand is represented by a minimum required average rate  $d_k$ , similar to that in [3], [4]. Generally speaking, the demand can be calculated from the QoS requirement of users at the test point. For example, at point  $k$ , we can assume that file transfer requests arrive following a Poisson point process with the mean arrival rate  $\lambda_k$  in  $\text{s}^{-1}$  and the exponentially distributed file size with mean  $L_k$  in bits, resulting in an average traffic load  $\Omega_k = \lambda_k L_k$  in bit/s [2]. The file transfer requests at the same test point are served according to a first-come-first-served policy. Hence, each test point has effectively an M/M/1 queue. Suppose the QoS of users at test point  $k$  requires the average file sojourn time (or response time) is not greater than a given value  $\tau_k$  (in second), which can be expressed as  $\frac{1}{R_k/L_k - \lambda_k} \leq \tau_k$  (see [25], [26]), where  $R_k$  is the average deliverable data rate of point  $k$  as given in (4), and  $R_k/L_k$  is the mean service rate in  $\text{s}^{-1}$ . This translates into the average data rate constraint  $R_k \geq d_k$ , as expressed in (8), with  $d_k = L_k/\tau_k + \Omega_k$ .

The proposed algorithms need the knowledge of user demand  $\{d_k, \forall k\}$  and deliverable rate  $\{r_{kbi}, \forall k, b, i\}$  at a central controller where resource management is executed. Location information is also required to identify which test point a user belongs to, which can be obtained by standard positioning methods. The serving BSs estimate/predict traffic pattern based on the traffic aggregation from all users within each test point [4]. Based on the estimated traffic information and QoS, the serving BS calculates the user demand of the associated test points and forwards this information to the central controller.

The deliverable rate  $r_{kbi}$  is calculated at the central controller. To facilitate this calculation, each BS should forward the channel gains between itself and all test points to the central controller. Note that channel information is routinely measured by the BSs in the current mobile network standards, either relying on uplink-downlink reciprocity or feedback from users. Since the resource management is adapted at a slow timescale, the BSs can only forward wideband channel coefficients  $G_{bk}$  in (2) and ignore the frequency-selective channel coefficients to reduce the overhead.

Once the central controller calculates the pattern resource parameter  $\pi$  and user resource parameter  $\alpha$  by proposed algorithms, it informs all BSs of these decision variables. Each BS is only allowed to have access to certain fraction of system bandwidth as specified by  $\pi_i$  (see Fig.1). Regarding the user association, the central controller decides which test point a user should belong to based on the location information. File transmission requests within a given test point  $k$  are routed to BS  $b$  by the central controller according to the association

TABLE VII

MAPPING JOINTLY OPTIMIZED SOLUTION TO CELL-SPECIFIC BIASES (IN DB). CASE I: 50 TEST POINTS,  $d_k = 0.2\text{Mbit/s}$ ; CASE II: 150 TEST POINTS,  $d_k = 0.2\text{Mbit/s}$ ; CASE III: 50 TEST POINTS,  $d_k = 1.0\text{Mbit/s}$ . ERROR IS DEFINED AS THE RATIO OF NUMBER OF WRONG ASSOCIATION TO THE NUMBER OF TEST POINTS.

Cell index	1,2,3	4	5	6	7	8	9	10	11	12	13	14	15	
Case I	0	0	40.7	0	25.2	4.8	40.6	42.6	0	32.1	0	0	38.3	0% error
Case II	0	0	41.0	36.8	25	35	40.7	40.6	0	34.8	33.1	33.4	35.9	1.33% error
Case III	0	28.1	40.9	36.4	31.1	33.1	40.6	42.6	0	32.1	35.7	5.6	38.3	0% error

decision. Then BS  $b$  allocates certain fraction of bandwidth under pattern  $i$  to serve the transmission, as specified by  $\alpha_{kbi}$ . On top of this adaptation, each BS can perform individual channel-aware scheduling for its associated users among the agreed spectrum in a more frequent manner to respond to fast fading channel fluctuations.

## VI. CONCLUSION

Interference coupling in heterogeneous networks introduces the inherent non-convexity to the multi-cell resource optimization problem, hindering the development of effective solutions. A new framework based on multi-pattern formulation has been proposed in this paper to study the energy efficient strategy for joint cell activation, user association and channel allocation. One key feature of this interference pattern formulation is that the patterns remain fixed and independent of the optimization process. This creates a favorable opportunity for convex formulation while still taking interference coupling into account. By grouping weakly mutual-interfering cells when formulating possible interference patterns in the network, and then allocating resources among these patterns, we arrive at an optimization problem with controllable complexity. A tailored algorithm has been proposed based on the reweighted  $\ell_1$  minimization and the cutting plane method in the dual domain by exploiting the problem structure, resulting in significant complexity saving. Relying on this algorithm, a benchmark involving all  $2^B$  possible patterns in the optimization has been derived to quantify the existing solutions with restricted patterns. Numerical results have demonstrated a high power saving by the proposed strategy. In contrast to previous studies on load balancing, per-tier biasing rule is not optimal for energy saving investigated in this paper.

## APPENDIX

### PROOF OF PROPOSITION 1

By letting  $\alpha_{kbi} = \pi_i \theta_{kbi}$ , the original problem can be equivalently rewritten as

$$\underset{\theta, \pi}{* \text{ minimize}} \quad P^{\text{tot}} = \sum_{b \in \mathcal{B}} [(1 - q_b) \rho_b P_b^{\text{OP}} + q_b |\rho_b|_0 P_b^{\text{OP}}] \quad (62a)$$

$$\text{subject to} \quad \rho_b = \sum_{i \in \mathcal{I}} \pi_i \sum_{k \in \mathcal{K}} \theta_{kbi}, \quad \forall b \quad (63)$$

$$\sum_{i \in \mathcal{I}} \pi_i \sum_{b \in \mathcal{B}} \theta_{kbi} r_{kbi} \geq d_k, \quad \forall k \quad (64)$$

$$\sum_{k \in \mathcal{K}} \theta_{kbi} \leq 1, \quad \forall b, \quad \forall i \quad (65)$$

$$\sum_{i \in \mathcal{I}} \pi_i = 1 \quad (66)$$

$$\pi_i \geq 0, \quad \forall i, \quad \theta_{kbi} \geq 0, \quad \forall k, b, i \quad (67)$$

In the following, we show that if an optimal solution  $(\theta^*, \pi^*)$  exists we can then obtain the same optimal objective with  $(\theta^*, \pi')$  where  $\pi'$  only has  $K + B + 1$  nonzero entries out of  $|\mathcal{I}|$  entries.

We first define  $\mathbf{t}_i = (t_{1i}, \dots, t_{bi}, \dots, t_{Bi})^T$  with  $t_{bi} = \sum_{k \in \mathcal{K}} \theta_{kbi}^*$ , and  $\mathbf{R}_i = (R_{1i}, \dots, R_{ki}, \dots, R_{Ki})^T$  with  $R_{ki} = \sum_{b \in \mathcal{B}} \theta_{kbi}^* r_{kbi}$ . Then define  $\boldsymbol{\rho} = (\rho_1, \dots, \rho_B)^T$  and  $\mathbf{d} = (d_1, \dots, d_K)^T$ . According to (63) and (64) (note that (64) must achieve equality at the optimum, otherwise the objective in (62) can be further reduced), the vector  $(\boldsymbol{\rho}^T, \mathbf{d}^T)^T = \sum_i \pi_i (\mathbf{t}_i^T, \mathbf{R}_i^T)^T$ , i.e., a convex combination of vectors  $(\mathbf{t}_i^T, \mathbf{R}_i^T)^T, \forall i \in \mathcal{I}$ , with  $\pi_i$  as coefficients. By Caratheodory's Theorem,  $(\boldsymbol{\rho}^T, \mathbf{d}^T)^T$  can be represented by at most  $K + B + 1$  of those vectors. Denoting the resulting coefficients by  $\pi'$ , we prove the proposition.

## REFERENCES

- [1] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 82–89, 2014.
- [2] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1525–1536, 2011.
- [3] E. Pollakis, R. Cavalcante, and S. Stanczak, "Base station selection for energy efficient network operation with the majorization-minimization algorithm," in *Signal Processing Advances in Wireless Communications (SPAWC), 2012 IEEE 13th International Workshop on*, 2012, pp. 219–223.
- [4] R. L. G. Cavalcante, S. Stańczak, M. Schubert, A. Eisenblätter, and U. Türke, "Toward Energy-Efficient 5G Wireless Communications Technologies: Tools for decoupling the scaling of networks from the growth of operating power," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 24–34, 2014.
- [5] S. Kim, S. Choi, and B. G. Lee, "A joint algorithm for base station operation and user association in heterogeneous networks," *IEEE Commun. Lett.*, vol. 17, no. 8, pp. 1552–1555, 2013.
- [6] L. Su, C. Yang, Z. Xu, and A. Molisch, "Energy-efficient downlink transmission with base station closing in small cell networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 4784–4788.
- [7] Y. Shi, J. Zhang, and K. Letaief, "Group sparse beamforming for green cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, 2014.
- [8] W.-C. Liao, M. Hong, Y.-F. Liu, and Z.-Q. Luo, "Base station activation and linear transceiver design for optimal resource management in heterogeneous networks," *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3939–3952, 2014.
- [9] Q. Kuang, W. Utschick, and A. Dotzler, "Optimal joint user association and resource allocation in heterogeneous networks via sparsity pursuit," *arXiv:1408.5091*, Aug. 2014.
- [10] B. Zhuang, D. Guo, and M. Honig, "Traffic-driven spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2027–2038, May 2015.
- [11] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted  $\ell_1$  minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5–6, pp. 877–905, 2008.

- [12] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, 2011.
- [13] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, 2013.
- [14] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 74–79, 2010.
- [15] Y. Nesterov, A. Nemirovskii, and Y. Ye, *Interior-point polynomial algorithms in convex programming*. SIAM, 1994, vol. 13.
- [16] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, New York, USA, 2004.
- [17] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear programming: theory and algorithms*, 3rd ed. New York: Wiley-Interscience, 2006.
- [18] Q. Kuang, J. Speidel, and H. Droste, "Joint base-station association, channel assignment, beamforming and power control in heterogeneous networks," in *Vehicular Technology Conference (VTC Spring), 2012 IEEE 75th*, 2012, pp. 1–5.
- [19] T. Lipp and S. Boyd, "Variations and extensions of the convex-concave procedure," 2014.
- [20] 3GPP, "Evolved universal terrestrial radio access (e-utra); physical layer; measurements (ts 36.214)," April 2011.
- [21] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous LTE-A cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 9, pp. 1479–1489, 2010.
- [22] 3GPP, "Further advancements for e-utra physical layer aspects (tr 36.814)," vol. v9.0.0, 2010.
- [23] A. Fehske, F. Richter, and G. Fettweis, "Energy efficiency improvements through micro sites in cellular mobile radio networks," in *GLOBECOM Workshops, 2009 IEEE*, 2009, pp. 1–5.
- [24] Q. Kuang, "Joint user association and reuse pattern selection in heterogeneous networks," in *Proc. IEEE Int. Symp. on Wireless Communication Systems (ISWCS)*, Spain, August 2014.
- [25] R. Nelson, *Probability, stochastic process, and queueing theory: the mathematics of computer performance modeling*, Springer, 1995.
- [26] B. Zhuang, D. Guo, and M. Honig, "Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, to appear, available: *arXiv:1509.04805*, Sep. 2015.
- [27] S. Luo, R. Zhang, and T. J. Lim, "Downlink and uplink energy minimization through user association and beamforming in C-RAN," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 494–508, 2015.
- [28] Q. Kuang, X. Yu, and W. Utschick, "Network topology adaptation and interference coordination for energy saving in heterogeneous networks", accepted at *ICASSP 2016*, available : *arXiv:1511.06888*.
- [29] G. R. Lanckriet, and B. K. Sriperumbudur, "On the convergence of the concave-convex procedure." *Advances in neural information processing systems*, pp. 1759 – 1767. 2009.
- [30] Gurobi Optimization, Inc., "Gurobi optimizer reference manual," <http://www.gurobi.com>, 2015.